Safe RLHF: Safe Reinforcement Learning from Human Feedback

Year: 2023 | Citations: 502 | Authors: Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu

---

Abstract

With the development of large language models (LLMs), striking a balance between the performance and safety of AI systems has never been more critical. However, the inherent tension between the objectives of helpfulness and harmlessness presents a significant challenge during LLM training. To address this issue, we propose Safe Reinforcement Learning from Human Feedback (Safe RLHF), a novel algorithm for human value alignment. Safe RLHF explicitly decouples human preferences regarding helpfulness and harmlessness, effectively avoiding the crowdworkers' confusion about the tension and allowing us to train separate reward and cost models. We formalize the safety concern of LLMs as an optimization task of maximizing the reward function while satisfying specified cost constraints. Leveraging the Lagrangian method to solve this constrained problem, Safe RLHF dynamically adjusts the balance between the two objectives during fine-tuning. Through a three-round fine-tuning using Safe RLHF, we demonstrate a superior ability to mitigate harmful responses while enhancing model performance compared to existing value-aligned algorithms. Experimentally, we fine-tuned the Alpaca-7B using Safe RLHF and aligned it with collected human preferences, significantly improving its helpfulness and harmlessness according to human evaluations.